Simulation Results (Thursday March 2nd)

Nirupama Tamvada

Data Generation Process

We generate the competing risks data from the cause-specific hazards following Beyersmann et al. (2009).

For both settings here, we consider the cause-specific hazards to be specified by a Weibull distribution with scale parameters 1.9 and 1.3 respectively for cause 1 and 2, which gives cause 1 (the main cause of interest) a higher density of incidence. The number of censored individuals $T*$ is specified by a $T* = Unif(1, 4)$ distribution which corresponds to approximately a 20 % rate of censoring.

Performance Measures

At the moment, we are only considering variable selection performance measures (i.e Sensitivity - number of true non-zero coefficients selected, Specificity: number of true zeroes called as zeroes, 1- Sensitivity: number of zero coefficients selected as non-zero and 1- Specificity: number of non-zero coefficients selected as zero).

Competing Models

1. Boosted Fine-Gray Model

- The Fine-Gray subdistribution hazard model has become the default method to estimate the incidence of outcomes over time in the presence of competing risks.
- To analyze competing risks data, the Cumulative Incidence Function (CIF) is calculated, which estimates the marginal probability for each competing event.

  Assuming that cause 1 is the primary cause of interest, the probability of experiencing cause 1 before time t for a given covariate set X is given by

$$P_1(t; X) = P(T \leq t, \epsilon = 1|X)$$

- Fine and Gray (1999) proposed a proportional hazards model to model the CIF with covariates, by treating the CIF curve as a subdistribution function.
- The subdistribution function is analogous to the Cox proportional hazard model, except that it models a hazard function (as known as subdistribution hazard) which was derived from a CIF
- The Fine-Gray sub-distribution hazard function estimates the hazard rate for event type c at time t based on the risk set that remains at time t after accounting for all previously occurring event types, which includes competing events.
- Disadvantages: For some subjects and for some time point patterns, it has been shown that the sum of the subject-specific probabilities for the risk of different event types (i.e the Cumulative Failure Probability) can exceed one which hampers the interpretability of the model in these cases.
- A commonly used implemented form of the Fine-Gray model is the boosted model, which fits this model with componentwise likelihood based boosting.
- This fit is especially suited for models with a large number of predictors as in our setting of $p > n$
- We optimize the line-search step-size modification factor using cross-validation: used default grid produced by `CoxBoost`

2. Penalized Binomial Model

- Direct modelling and assessment of covariate effects for the cumulative incidence curve via binomial regression: quite similar to what casebase aims to do (Scheike et al. (2008))
- The following model is considered:

$$g\{P_1(t;X)\} = \eta(t) + X^T \gamma_x$$

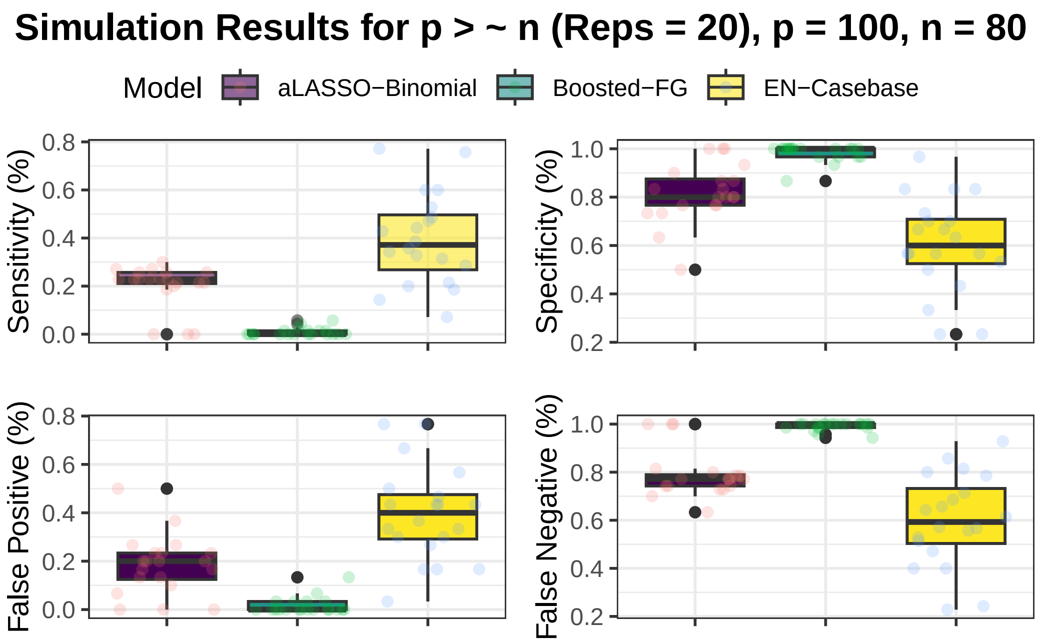
  where $\eta(t)$ represents the effects of X_i on the cumulative incidence curve at time t and g represents a link function (the logit here)
- The parameters of the model are estimated and censoring is accounted for by using inverse probability of censoring weighting.
- We use adaptive LASSO as the regularization penalty as this was implemented in the original paper introducing the penalized version of this model (Ambrogi et al. (2016))

  . The weights for the penalized variables come from the coefficients from a ridge regression. Cross-validation was performed to tune λ using `glmnet`
- This is probably the most important competitor model for `casebase`.

1. p $>\approx$ n setting (N = 80, p = 100)

We consider the case where p is greater than n but only marginally so. Here, we consider a mildly sparse setting with the sparsity of β_1 at 70 non-zero coefficients, whereas β_2 has 90 non-zero coefficients. The set $[1.5, -1, 1.5, 1]$ contains the possible magnitude of the coefficient values. We generate coefficients from a multivariate normal distribution with mean 0 and an AR(1) correlation matrix with $Cor(X_i, X_j) = \rho^{|i-j|}$ where ρ was set to 0.5.

```
beta1 <- c(rep(-1.5, 5), rep(1, 30),
    rep(1.5, 5), rep(-1.5, 5), rep(0, 30), rep(1.5, 5), rep(1.5, 20))
beta2 <- c(rep(-1.5, 5), rep(0, 10),
rep(1.5, 5), rep(-1, 5), rep(1.5, 5), rep(1.5, 5),
rep(-1.5, 5),rep(1, 30),rep(1, 10), rep(1.5, 5), rep(1, 15))
```



Notes

- The last column of the covariates (time) is not penalized in the case-base model, sensible?
- Optimize grid-search for casebase? - other performance measures
  - less correlations, block structure only one
  - vanilla simulation -

- elastic net - tune alpha; higher alpha in elastic net
- - reproduce simulation results in Binomial model paper

- Group LASSO? - nah not necessary
- `CoxBoost` might need some specification of step-size values for grid instead of default or mybe performance will improve in true p > n case

To do for next week

1. Figure out absolute risk prediction curves. Most papers seem to look at variable selection and then plot .632+ prediction error curves for absolute risk.
2. Add some complexity to data generation mechanism: 1. Probably easy to add measurement error noise to X (see Liu et al. 2021) 2. Figure out how to add outliers to time: can either generate from mixture of hazards for more realistic baseline hazard functions (hard to integrate but `uniroot` can probably find solution), add some white noise (`rnorm`) or mix some time-points from another Weibull hazard.
3. Replicate simulation study in Tapak et al., 2015